Universitetet i Bergen
Molekylærbiologisk institutt

Matematisk-naturvitenskapelig Embetseksamen

MOL204 Anvendt bioinformatikk I

bokmål / nynorsk / english

Torsdag 4. Mai 2006, 4 timer, kl 9:00-13:00

Alle spørsmål skal besvares. Dersom ikke annet er angitt, forventes korte og konsise svar. Les oppgavetekstene nøye. Sammenlign gjerne med den engelske teksten. Bruk ikke for lang tid på noe enkeltspørsmål. Hopp heller over spørsmål som synes å ta for lang tid, og ta heller disse til slutt. Dersom et spørsmål er uklart, gjør da oppmerksom på hvordan du har bestemt deg for å forstå det. Ta med alle relevante momenter i hver oppgave eller henvis til andre svar dersom noen av oppgavene synes å overlappe. I noen av spørsmålene er det brukt engelske ord slik de forekommer i læreboken. Dersom du illustrerer dine svar med figurer, husk at disse skal vurderes både som original og kopi - unngå derfor bruk av farger. Bruk heller andre metoder for å fremheve ulike deler av figurene. Tentative poeng er angitt for hver oppgave. Totalt utgjør de **78 poeng.** Bruk poengene til å vurdere hvor mye arbeid det lønner seg å legge i hvert svar. ***Merk: Ingen spørsmål krever lange utredninger.***

Tillatte hjelpemidler: Norsk tekst side 2-4.
kalkulator
ordbøker for språk

MOL204 Applied Bioinformatics I

Thursday May 4th. 2006, 4 hours, 9:00-13:00

Answer all questions. If not otherwise stated, brief and concise answers are expected. Read the questions carefully. Do not spend too much time on each question. It is better to proceed to the next questions and return to time consuming questions at the end. If a question appears unclear or ambiguous, then explain how you have interpreted the question. Include all relevant aspects in each answer. If one answer appears to overlap with another, you may cross reference them. If you illustrate any of your answers with figures, remember that the these will be evaluated both from the original and from the copy - avoid therefore the use of colors. Use other methods to highlight different parts of the figures. For each question is given a number of points to indicate how the question contributes to the total of **78 points.** Use these points to judge how much time it is worth spending on each question. ***Note: None of the questions require long answers***

Allowed aids: English text pages 5-7
electronic calculator
language dictionaries

Oppgave 1 – Global sammenstilling av sekvenser (totalt 11p)

Vi har gitt to sekvenser `SVLSF` og `SFF` og følgende scoringsmatrise (utdrag fra PAM 250):

```
    F  L  S  V
F   9  2 -3 -1
L      6 -3  2
S         2 -1
V            4
```

Vi bruker en lineær gapkostnad med kostnad 2 for hvert gap. Matrisen *H* for å finne beste globale sammenstilling ved dynamisk programmering ser slik ut, delvis utfylt:

```
            S   V   L   S   F
      0    -2  -4  -6  -8 -10
S    -2     2   0  -2  -4  -6
F    -4     0   1   2
F    -6    -2  -1   3
```

A (4p) Fyll ut resten av verdiene i matrisen *H*. Hva er score for beste sammenstilling?

B (4p) Finn den eller de beste globale sammenstillingene. Forklar fremgangsmåten, og illustrer med å tegne en eller flere stier gjennom matrisen.

C (3p) Sett at vi bytter ut den lineære gapkostnad med en *affin* gapkostnad, med kostnad 10 for å åpne et gap (gap av lengde 1) og 1 for hver utvidelse. Vi kan ikke lenger bruke algoritmen i punkt A, men hvordan tror du beste sammenstilling nå vil se ut (her må du resonnere deg fram til svaret uten å beregne den nye matrisen)? Hvilken score har denne sammenstillingen? Gi en biologisk begrunnelse for å velge en relativt høy kostnad for å åpne et gap.

Oppgave 2 – Scoringsmatriser (total 11p)

A (5p) Gi en kort beskrivelse av hovedtrinnene i Dayhoffs prosedyre for å konstruere PAM-scoringsmatriser.

B (3p) Hva er en *substitusjonsmatrise* (eller mutasjonssannsynlighetsmatrise)? En substitusjonsmatrise er ikke symmetrisk, i motsetning til en scoringsmatrise. Hvordan kan det tolkes?

C (3p) Hva er de viktigste forskjellene mellom PAM- og BLOSUM-matrisene?

Oppgave 3 - Multippel sekvenssammenstilling (totalt 13p)

A (4p) Hva er det man vil oppnå når man gjør en multippel sammenstilling av et sett med homologe proteinsekvenser? Forklar kort hvordan konserverte blokker (områder) i den multiple sammenstillingen relaterer til proteinenes struktur.

B (5p) Progressiv multippel sammenstilling:
(i) Forklar kort hovedtrinnene i fremgangsmåten.
(ii) Hva er svakheten med denne fremgangsmåten?

C (4p) I en av kursets obligatoriske øvelser så vi at vi fikk langt bedre multippel sammenstilling av proteinsekvenser med Clustal når vi brukte en strukturmaske. Beskriv kort hvordan strukturmasken brukes i Clustal og forklar hvorfor dette gir bedre sammenstillinger. Hvor kan du finne informasjon for å lage en strukturmaske?

Oppgave 4 - Sekvenssammenstilling og proteindomener (totalt 19p)

A (4p) I dette kurset har vi lagt stor vekt på at de fleste proteiner har en modulær oppbygning hvor ett eller flere globulære domener er knyttet sammen med mer eller mindre ustrukturerte sekvenssegmenter. Beskriv hva konserverte globulære domener er og forklar hvorfor de har en slik sentral betydning for sekvensanalyse av proteiner.

B (3p) Hva er Pfam og SMART? Hvorfor er søk i Pfam og/eller SMART så nyttige ved analyse av proteinsekvenser?

C (3p) I databasene SCOP og CATH er proteindomener er klassifisert på lignenende vis. Forklar kort hva som menes med familie, superfamilie og fold (Det er ikke nødvendig å forklare forskjellene mellom de to klassifiseringssystemene).

D (4p) Det sies at proteiner med samme fold ikke nødvendigvis er homologe. Kommenter dette utsagnet og forklar hva årsaken til dette kan være. Tror du at du vil kunne finne ikke-homologe proteiner med samme fold ved hjelp av vanlige Blastp-søk? Begrunn svaret.

E (3p) Forklar kort hva programmet PSI-BLAST er og hvorfor det er spesielt godt egnet til å finne nye og ukjente domener i en sekvensfamilie.

F (2p) Forklar kort hvordan resiproke søk kan brukes til å vurdere om en matchende sekvens fra et databasesøk med PSI-BLAST er en sann positive eller ikke. Begrunn svaret.

Oppgave 5 - Informasjon i sekvensdatabaser (totalt 12p)

A (1p) Hvorfor er annotasjon av proteiner i SwissProt regnet å ha bedre kvalitet enn annotasjon i TrEMBL (translasjon av EMBL-databasen)?

B (4p) Forklar kort hva som menes med "features" i SwissProt og nevn 3 typer annotasjon som kan forekomme i en SwissProt "Feature-tabell".

C (5p) Det sies at det er en del feil annotasjon i sekvensdatabasene. Nevn to mulige kilder til feil i annotasjon. For hver av de to feilkildene, forklar hvordan du selv kan undersøke om det er sannsynlig at annotasjonen er korrekt eller ikke. Svarene må begrunnes.

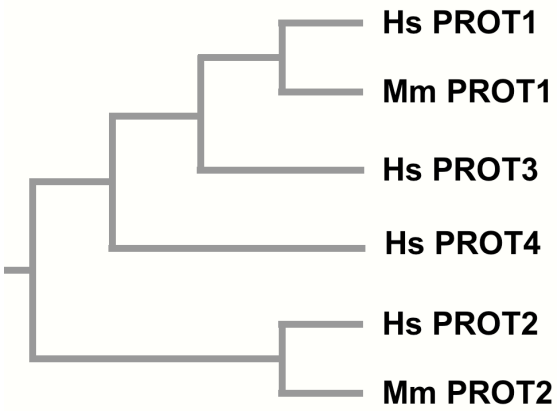
D (2p) Hvorfor er ikke SwissProt egnet til å søke blant alle proteiner i det humane proteom? Nevn en annen sekvensdatabase som er bedre egnet til dette formål.

Oppgave 6 - Trær (totalt 12p)

A (4p) Forklar forskjellen mellom karakterbaserte og avstandsbaserte metoder for estimering av fylogenetiske trær og nevn minst to eksempler på karakter-baserte metoder.

B (3p) Forklar kort hva som menes med et rotet og et urotet fylogenetisk tre. Gitt at du har en metode for å konstruere urotede trær, forklar hvordan en utgruppe (engelsk: *outgroup*) kan brukes til å plassere roten i et slikt urotet tre.

C (3p) Figuren nedenfor viser et rotet tre av homologe proteiner fra mus og menneske. Hvilke proteiner er paraloge og hvilke er ortologe til Hs PROT1 ? ('Hs' betyr *Homo sapiens,* 'Mm' betyr *Mus musculus*).



D (2p) Basert på treet ovenfor, forklar kort hvordan man kan tenke seg at denne proteinfamilien har utviklet seg gjennom evolusjonen.

end of Norwegian text - English text on next pages

English text

Question 1 - Global alignment of sequences (total 11p)

We are given two sequences `SVLSF` and `SFF` and the following scoring matrix (excerpt from PAM 250):

```
    F  L  S  V
F   9  2 -3 -1
L      6 -3  2
S         2 -1
V            4
```

We use a linear gap cost with cost 2 for each gap. The matrix H used to find the best global alignment by dynamic programming is shown below, partially filled in:

```
            S   V   L   S   F
      0   -2  -4  -6  -8 -10
S   -2    2   0  -2  -4  -6
F   -4    0   1   2
F   -6   -2  -1   3
```

A (4p) Fill in the remaining values in matrix *H*. What is the score for the best alignment?

B (4p) Find the best global alignment(s). Explain the procedure and illustrate by a drawing one or more paths through the matrix.

C (3p) Assume that we replace the linear gap cost with an *affine* gap cost, with a cost of 10 for opening a gap (gap of length 1) and 1 for extending it. We can then no longer use the algorithm in A, but what do you think the best alignment will be (you must use reasoning to find this answer without calculating the new matrix)? Which score does this alignment have? Give a biological explanation for choosing a relatively high cost for opening a gap.

Question 2 - Scoring matrices (total 11p)

A (5p) Give a brief description the main steps in Dayhoff's procedure for generating PAM scoring matrices.

B (3p) What is a *substitution matrix* (or mutation probability matrix)? A substitution matrix is not symmetrical as opposed to a scoring matrix. How can this be interpreted?

C (3p) What are the most important differences between PAM- and BLOSUM-matrices?

Question 3 - Multiple sequence alignment (total 13p)

A (4p) What is the goal when generating a multiple sequence alignment of a set of homologous protein sequences? Explain briefly how conserved blocks (regions) in the multiple alignment relates to the structures of the proteins.

B (5p) Progressive multiple alignment:
(i) Explain briefly the main steps for this procedure.
(ii) What is the weakness with this procedure?

C (4p) In one of the exercises in the course, we obtained far better multiple alignments of protein sequences with Clustal when we used a structure mask. Explain briefly how the structure mask is used in Clustal and explain why it gives better alignments. Where can you find information which you can use to build a structure mask?

Question 4 - Sequence alignment and protein domains (total 19p)

A (4p) In this course we have emphasized that most proteins have a modular architecture where one or more globular domains are linked with more or less unstructured sequence segments. Describe what conserved globular domains are and explain why they have such a central role in sequence analysis of proteins.

B (3p) What is Pfam and SMART? Why are searches with Pfam and/or SMART so useful for analysis of protein sequences?

C (3p) In the databases SCOP and CATH, protein domains are classified in a similar way. Explain briefly what is meant by family, superfamily and fold (it is not necessary to explain the differences between the two classification schemes).

D (4p) It is said that proteins with the same fold may not be homologous. Comment this statement and explain the reason why this may be so. Do you think you can find non-homologous sequences with the same fold using ordinary Blastp searches? Justify your answer.

E (3p) Explain briefly what PSI-BLAST is and why it is so well suited to identify new and unknown domains in a protein family.

F (2p) Explain briefly how reciprocal searches can be used to evaluate whether a matching sequence from a database search with PSI-BLAST is a true positive or not. Justify your answer.

Question 5 - Information in sequence databases (total 12p)

A (1p) Why is annotation of proteins in SwissProt considered to have better quality than annotation in TrEMBL (translation of the EMBL database)?

B (4p) Explain briefly what is meant by features in SwissProt and mention 3 types of annotation that can appear in a SwissProt Feature table.

C (5p) It is often said that there can be errors in the annotation in the sequence databases. Mention two possible sources of errors in annotation. For each of the two sources of error, explain how you could investigate whether it is likely that the annotation is correct or not. Justify your answers.

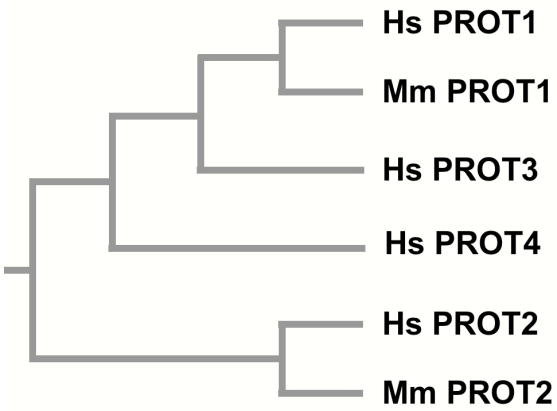
D (2p) Why is SwissProt not a suitable database for searching among all proteins in the human proteome? Mention another sequence database that would be better suited for this purpose.

Question 6 - Trees (total 12p)

A (4p) Explain the difference between character-based and distance-based methods for estimating phylogenetic trees and mention at least two examples of character-based methods.

B (3p) Explain briefly what is meant by a rooted and an unrooted phylogenetic tree. If you have a method for constructing unrooted trees, explain how an outgroup could be used to place the root on such an unrooted tree.

C (3p) The figure below shows a rooted tree of homologous proteins from mouse and man. Which proteins are paralogs and which are orthologs of Hs PROT1? ('Hs' means *Homo sapiens,* 'Mm' means *Mus musculus*).



D (2p) On the basis of the tree shown above, explain briefly how this protein family may have evolved.

end of English text